Dimensionality Reduction and Classification Feature Using Mutual Information Applied to Hyperspectral Images: A Wrapper Strategy Algorithm Based on Minimizing the Error Probability Using the Inequality of Fano

ELkebir Sarhrouni*, Ahmed Hammouch** and Driss Aboutajdine*

*LRIT, Faculty of Sciences, Mohamed V - Agdal University, Morocco
**LRGE, ENSET, Mohamed V - Souissi University, Morocco
sarhrouni436@yahoo.fr

Abstract

Hyperspectral image is a substitution of more than a hundred images, called bands, of the same region. They are taken at juxtaposed frequencies. The reference image of the region is called Ground Truth map (GT). the problematic is how to find the good bands to classify the pixels of regions; because the bands can be not only redundant, but a source of confusion, and decreasing so the accuracy of classification. Some methods use Mutual Information (MI) and threshold, to select relevant bands without treatement of redundancy; others consider the neighbors having sensibly the same MI with the GT as redundant and so discarded. This is the most inconvenient of this method, because this avoids the advantage of hyperspectral images: some precious information can be discarded. In this paper well make difference between useful and useless redundancy. A band contains useful redundancy if it contributes to decreasing error probability. According to this scheme, we introduce new algorithm using also mutual information, but it retains only the bands minimizing the error probability of classification. To control redundancy, we introduce a complementary threshold. So the good band candidate must contribute to decrease the last error probability augmented by the threshold. This process is a wrapper strategy; it gets high performance of classification accuracy but it is expensive than filter strategy.

Keywords: Hyperspectral images, classification, feature selection, mutual information, error probability, redundancy

1   Introduction

In the feature classification domain, the choice of data affects widely the results. For the Hyperspectral image, the bands dont all contain the information; some bands are irrelevant like those affected by various atmospheric effects, see Figure.4, and decrease the classification accuracy. And there exist redundant bands to complicate the learning system and product incorrect prediction [14]. Even the bands contain enough information about the scene they may can't predict the classes correctly if the dimension of space images, see Figure.3, is so large that needs many cases to detect the relationship between the bands and the scene (Hughes phenomenon) [10]. We can reduce the dimensionality of hyperspectral images by selecting only the relevant bands (feature selection or subset selection methodology), or extracting, from the original bands, new bands containing the maximal information about the classes, using any functions, logical or numerical (feature extraction methodology) [11][9]. Here we focus on the feature selection using mutual information. Hyperspectral images have three advantages regarding the multispectral images [6],

Assertion: *when we reduce hyperspectral images dimensionality, any method used must save the precision and high discrimination of substances given by hyperspectral image.*
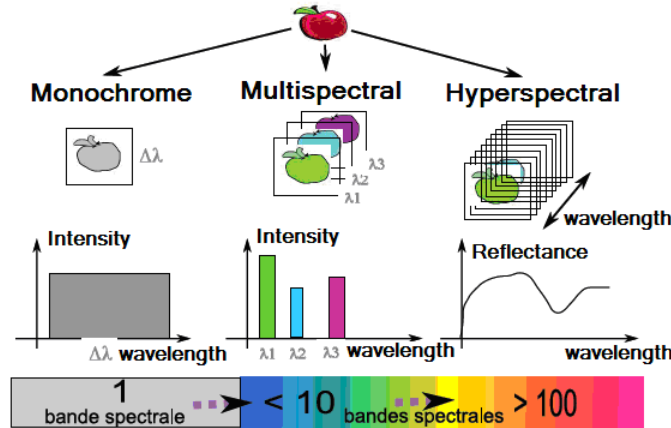

Figure 1: Precision an dicrimination added by hyperspectral images

In this paper we use the Hyperspectral image AVIRIS 92AV3C (Airborne Visible Infrared Imaging Spectrometer). [2]. It contains 220 images taken on the region Indiana Pine at north-western Indiana, USA [1]. The 220 called bands are taken between 0.4m and 2.5m. Each band has 145 lines and 145 columns. The ground truth map is also provided, but only 10366 pixels are labeled fro 1 to 16. Each label indicates one from 16 classes. The zeros indicate pixels how are not classified yet, Figure.2.

The hyperspectral image AVIRIS 92AV3C contains numbers between 955 and 9406. Each pixel of the ground truth map has a set of 220 numbers (measures) along the hyperspectral image. This numbers (measures) represent the reflectance of the pixel in each band. So the pixel is shown as vector off 220 components. Figure .3.
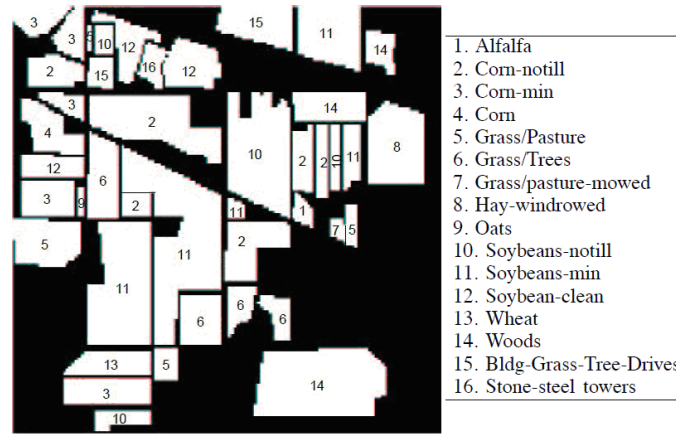

Figure 2: The Ground Truth map of AVIRIS 92AV3C and the 16 classes

The hyperspectral image AVIRIS 92AV3C contains numbers between 955 and 9406. Each pixel of the ground truth map has a set of 220 numbers (measures) along the hyperspectral image. This numbers (measures) represent the reflectance of the pixel in each band. So the pixel is shown as vector off 220 components.
Figure.3 shows the vector pixels notion [7]. So reducing dimensionality means selecting only the dimensions caring a lot of information regarding the classes.
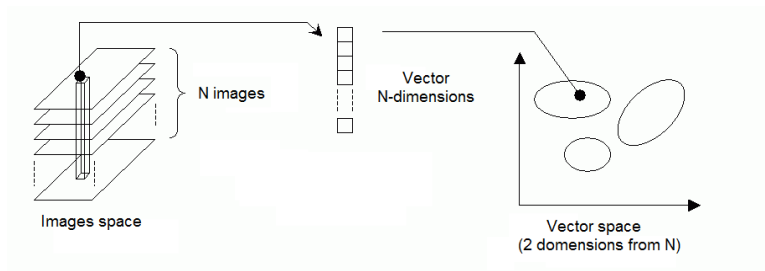

Figure 3: The notion of pixel vector

We can also note that not all classes are carrier of information. In Figure. 4, for example, we can show the effects of atmospheric affects on bands: 155, 220 and other bands. This hyperspectral image presents the problematic of dimensionality reduction.

2   Mutual Information based feature selection

2.1   Definition of mutual information

This is a measure of exchanged information between tow ensembles of random variables A and B :

$$I(A, B) = \sum log_2 \; p(A, B) \; \frac{p(A, B)}{p(A).p(B)}$$

Considering the ground truth map, and bands as ensembles of random variables, we calculate their interdependence.

Geo [3] uses also the average of bands 170 to 210, to product an estimated ground truth map, and use it instead of the real truth map. Their curves are similar. This is shown at Figure 4.
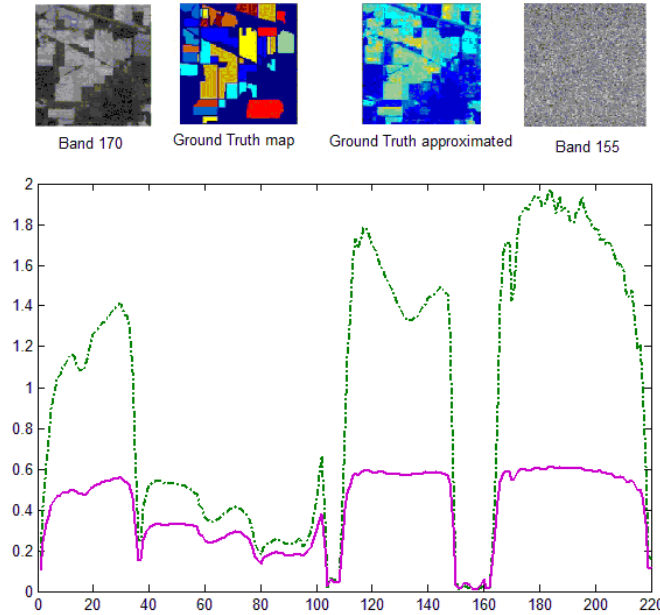

Figure 4: Mutual information of AVIRIS with the Ground Truth map (solid line) and with the ground apporoximated by averaging bands 170 to 210 (dashed line) .

2.2   The mesure of error probability

Fano [14] has demonstrated that as soon as mutual information of already selected feature has high value, the error probability of classification is de-

creasing, according to the formula bellow:

$$\frac{H(C/X)-1}{Log_2(N_c)} \leq P_e \leq \frac{H(C/X)}{Log_2}$$

with :

$$\frac{H(C/X)-1}{Log_2(N_c)} = \frac{H(C)-I(C;X)-1}{Log_2(N_c)}$$

and :

$$P_e \leq \frac{H(C)-I(C;X)}{Log_2} = \frac{H(C/X)}{Log_2}$$

The expression of conditional entropy *H(C/X)* is calculated between the ground truth map (i.e. the classes C) and the subset of bands candidates X. Nc is the number of classes. So when the features X have a higher value of mutual information with the ground truth map, (is more near to the ground truth map), the error probability will be lower. But its difficult to compute this conjoint mutual information *I(C;X)*, regarding the high dimensionality [14].This propriety makes Mutual Information a good criterion to measure resemblance between too bands, like its exploited in section II. Furthermore, we will interest at case of one feature candidate X.

Corollary: for one feature X, as X approaches the ground truth map, the interval P_e is very small.

3 The principe of proposed algorithm based on inequality of Fano

Our idea is based on this observation: the band that has higher value of Mutual Information with the ground truth map can be a good approximation of it. So we note that the subset of selected bands are the good ones, if thy can generate an estimated reference map, sensibly equal the ground truth map. Its clearly thats an Incremental Wrapper-based Subset Selection (IWSS) approach[16] [13].

Our process of band selection will be as following: we order the bands according to value of its mutual information with the ground truth map. Then we initialize the selected bands ensemble with the band that has highest value of MI. At a point of process, we build an approximated reference map C_{est} with the already selected bands, and we put it instead of X for computing the error probability (P_e); the latest band added (at those already selected) must make P_e decreasing, if else it will be discarded from the ensemble retained. Then we introduce a complementary threshold T_h to control redundancy. So the band to be selected must make error probability less than (P_e - T_h) , where P_e is calculated before adding it. The algorithm following shows more

detail of the process:

Let SS be the ensemble of bands already selected and S the band candidate to be selected. $Build_{estimated}C()$ is a procedure to construct the estimated reference map. P_e is initialized with a value P_e^* . X the number of bands to be selected, SS is empty and $R = 1..220$.

Algorithm 1 Proposed for Dimentionality Reduction and Redundancy control

---

   while $|SS| < X$ **do**

      Select *band index$_s$* $S = argmax_s$ MI(s)

      $SS \leftarrow SS \cup S$ and $R \leftarrow R \setminus S$

      $C_{est} = Build_{estimated}C(SS)$

$$Pe = \frac{H(C/C_{est})}{Log_2} - \frac{H(C/C_{est}) - 1}{Log_2(N_c)};$$

      if $Pe \leq Pe^* - Th$ **then**

         $Pe \leftarrow Pe^*$

      else

         $SS \leftarrow SS \setminus S$

      end if

   end while

---

4   Results and analysis

We apply this algorithm on the hyperspectral image AVIRIS 92AV3C [1], 50% of the labeled pixels are randomly chosen and used in training; and the other 50% are used for testing classifcation [3]. The classifer used is the SVM [5] [12] [4].

The procedure to construct the estimated reference map C_{est} is the same SVM classifier used for classification. So C_{est} is the output of classification.

4.1   Results

Table. I shows the results obtained for several thresholds. We can see the effectiveness selection bands of our algorithm, and the important effect of avoiding redundancy.

Table 1: Results illustrate elimination of Redundancy using algorithm based on inequality of Fano, for thresholds (Th)

$Bands$ $retained$	Th					
	0.00	0.001	0.008	0.015	0.020	0.030
10	55.43	55.43	55.58	53.09	60.06	71.62
18	59.09	59.09	64.41	73.70	82.62	90.00
20	63.08	63.08	68.50	76.15	84.36	-
25	66.02	66.12	74.62	84.41	89.06	-
27	69.47	69.47	76.00	86.73	91.70	-
30	73.54	73.54	79.04	88.68	-	-
35	76.06	76.06	81.38	92.36	-	-
40	78.96	79.41	86.48	-	-	-
45	80.58	80.60	89.09	-	-	-
50	81.63	81.20	91.14	-	-	-
53	82.27	81.22	92.67	-	-	-
60	86.13	86.23	-	-	-	-
70	86.97	87.55	-	-	-	-
80	89.11	89.42	-	-	-	-
90	90.55	90.92	-	-	-	-
100	92.50	93.18	-	-	-	-
102	92.62	93.34	-	-	-	-
114	93.34	-	-	-	-	-

Figure.5 shows more detail of the accuracy curves, versus number of bands retained, for several thresholds. This covers all behaviors of the algorithm.

4.2 Analysis

Table.I and Figure.5 allow us to comment four cases:

First: For the highest threshold values (0.1, 0.05, 0.03 and 0.02) we obtain a hard selection: a few more informative bands are selected; the accuracy of classification is 90% with less than 20 bands selected.

Second: For the medium threshold values (0.015, 0.012, 0.010, 0.008, 0.006), some redundancy is allowed, in order to made increasing the classification accuracy.

Tired: For the small threshold values (0.001 and 0) , the redundancy allowed becomes useless, we have the same accuracy with more bands.

Finally: for the negative thresholds, for example -0.01, we allow all bands to be selected, and we have no action of the algorithm. This corresponds at selection bay ordering bands on mutual information . The performance is low.
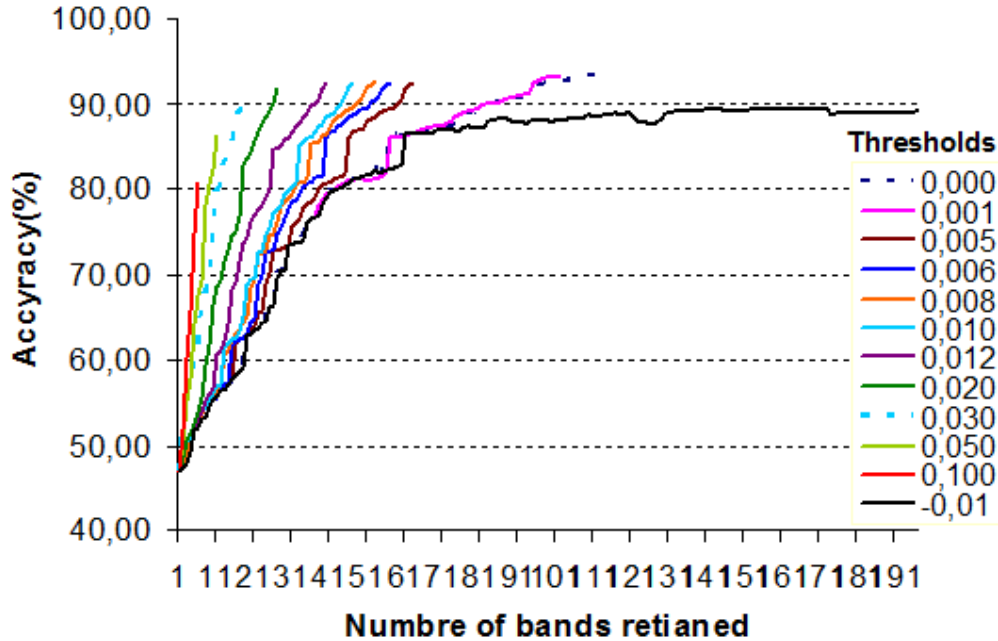
Figure 5: Accuracy of classification using the algorithm based on inequality of Fano, using numerous thresholds.

We can not here that [15] uses two axioms to characterize feature selection. Sufficiency axiom: the subset selected feature must be able to reproduce the training simples without losing information. The necessity axiom "simplest among different alternatives is preferred for prediction". In the algorithm proposed, reducing error probability between the truth map and the estimated minimize the information loosed for the samples training and also the predicate ones.

We note also that we can use the number of features selected like condition to stop the search; so we can obtain an hybrid approach filter-wrapper[16].

Partial conclusion:
The algorithm proposed is a very good method to reduce dimensionality of hyperspectral images.

We illustrate in Figure .6, the Ground Truth map originally displayed, like at Figure .1, and the scene classified with our method, for threshold as 0.03, so 18 bands selected.
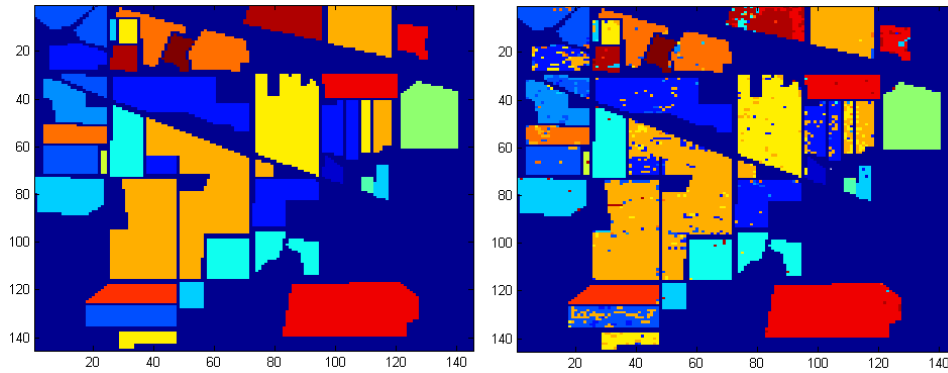
Figure 6: Original Grand Truth map(in the left) and the map produced bay our algorithm according to the threshold 0.03 i.e 18 bands (in the right). Acuracy=90%.

Table II indicates the classification accuracy of each class, for several thresholds.

Table 2: Accuracy of classification(%) of each class for numerous thresholds (Th)

$Class$	$Total$				Th		
	p$ixels$	0.00	0.001	0.008	0.015	0.020	0.030
1 :	54	86.96	82.61	86.96	83.96	78.26	86.96
2 :	1434	91.07	89.40	89.54	89.12	88.01	83.96
3 :	834	89.93	90.89	89.69	86.09	83.69	81.53
4 :	234	96.32	83.76	86.32	87.18	87.18	86.32
5 :	597	95.93	95.53	94.34	95.93	95.93	95.53
6 :	747	98.60	98.60	98.32	98.60	98.32	98.32
7 :	26	84.62	84.62	84.62	84.62	84.62	84.62
8 :	489	98.37	98.37	98.78	97.96	98.78	98.78
9 :	20	100	100	100	100	100	100
10:	968	92.15	92.98	91.32	91.74	90.91	89.05
11:	2468	93.84	94.17	92.54	92.71	91.90	91.25
12:	614	91.21	93.49	92.83	92.18	88.93	87.30
13:	212	98.06	98.06	98.06	98.06	98.06	98.06
14:	1294	97.53	97.86	97.22	97.84	97.99	97.53
15:	390	79.52	77.71	75.90	74.10	78.92	64.46
16:	95	93.48	93.48	93.48	93.48	93.48	93.48

Comments:
First :we can not the effectiveness of this algorithm for particularly the classes with a few number of pixels, for example class number 9.
Second: we can not that 18 bands (i.e. threshold 0.03) are sufficient to detect materials contained in the region. Its also shown in Figure .6
Tired: one of important comment is that most of class accuracy change lately when the threshold changes between 0.03 and 0.015

5   Conclusion

In this paper we presented the necessity to reduce the number of bands, in classification of Hyperspectral images. Then we have introduce the mutual information based scheme. We carried out their effectiveness to select bands able to classify the pixels of ground truth. We introduce an algorithm also based on mutual information and using a measure of error probability (inequality of Fano). To choice a band, it must contributes to reduce error probability. A complementary threshold is added to avoid redundancy. So each band retained has to contribute to reduce error probability by a step equal to threshold even if it caries a redundant information. We can tell that we conserve the useful redundancy by adjusting the complementary threshold. The process introduced is able to select the good bands to classification for also the classes that have a few number of pixels. This algorithm is a feature selection methodology. But its a wrapper approach, because we use the classifier to make the estimated reference map. This is expenssive than Filter strategy, but it can be used for application that need more precision. This scheme is very interesting to investigate and improve, considering its performance.

References

[1] D. Landgrebe, On information extraction principles for hyperspectral data: A white paper, Purdue University, West Lafayette, IN, Technical Report, School of Electrical and Computer Engineering, 1997. Tlchargeable ici : http://dynamo.ecn.purdue.edu/∼landgreb/whitepaper.pdf.

[2] ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/

[3] Baofeng Guo, Steve R. Gunn, R. I. Damper Senior Member, "Band Selection for Hyperspectral Image Classification Using Mutual Information" , IEEE and J. D. B. Nelson. IEEE GEOSCIENCE AND REMOTE SINSING LETTERS, Vol .3, NO .4, OCTOBER 2006.

[4] Baofeng Guo, Steve R. Gunn, R. I. Damper, Senior Member, IEEE, and James D. B. Nelson."Customizing Kernel Functions for SVM-Based Hyperspectral Image Classification", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 17, NO. 4, APRIL 2008.

[5] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology , 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] Nathalie GORRETTA-MONTEIRO , Proposition d'une approche de'segmentation dimages hyperspectrales. PhD thesis. Universite Montpellier II. Fvrier 2009.

[7] David Kernis." Amlioration de la classification automatique des fonds marins par la fusion multicapteurs acoustiques". Thse, ENST BRETAGNE, universit de Rennes. Chapitre3, Rduction de dimensionalit et classification,Page.48. Avril 2007.

[8] Kwak, N and Choi, C. "Featutre extraction based on direct calculation of mutual information".IJPRAI VOL. 21, NO. 7, PP. 1213-1231, NOV. 2007 (2007).

[9] Nojun Kwak and C. Kim,"Dimensionality Reduction Based on ICA Regression Problem". ARTIFICIAL NEURAL NETWORKS-ICANN 2006. Lecture Notes in Computer Science, 2006, Isbn 978-3-540-38625-4, Volume 1431/2006.

[10] Huges, G. Information Thaory,"On the mean accuracy of statistical pattern recognizers". IEEE Transactionon Jan 1968, Volume 14, Issue:1, p:55-63, ISSN 0018-9448 DOI: 10.1109/TIT.1968.1054102.

[11] YANG, Yiming, and Jan O. PEDERSEN, 1997.A comparative study of feature selection in text categorization. In: ICML. 97: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA:Morgan Kaufmann Publishers Inc., pp. 412.420.

[12] Chih-Wei Hsu; Chih-Jen Lin,"A comparison of methods for multiclass support vector machines" ;Dept. of Comput. Sci. Inf. Eng., Nat Taiwan Univ. Taipei Mar 2002, Volume: 13 I:2;pages: 415 - 425 ISSN: 1045-9227, IAN: 7224559, DOI: 10.1109/72.991427

[13] Bermejo, P.; Gamez, J.A.; Puerta, J.M."Incremental Wrapper-based subset Selection with replacement: An advantageous alternative to sequential forward selection" ;Comput. Syst. Dept., Univ. de Castilla-La Mancha,

Albacete; Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on March 30 2009-April 2 2009, pages: 367 - 374, ISBN: 978-1-4244-2765-9, IAN: 10647089, DOI: 10.1109/CIDM.2009.4938673

[14] Lei Yu, Huan Liu,"Efficient Feature Selection via Analysis of Relevance and Redundancy", Department of Computer Science and Engineering; Arizona State University, Tempe, AZ 85287-8809, USA, Journal of Machine Learning Research 5 (2004) 1205-1224.

[15] Hui Wang, David Bell, and Fionn Murtagh, "Feature subset selection based on relevance" , Vistas in Astronomy, Volume 41, Issue 3, 1997, Pages 387-396.

[16] P. Bermejo, J.A. Gmez, and J.M. Puerta, "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-